Adversarial ML Security Assessment Report

Generated: 2026-01-10 14:50:01

EXECUTIVE SUMMARY

Project Overview

This report summarizes the security assessment of the MNIST CNN model against various adversarial attacks and evaluates the effectiveness of multiple defense mechanisms.

Key Findings

Risk Assessment

Recommendations

1. Implement adversarial training for critical deployments 2. Use input smoothing as a lightweight defense 3. Deploy ensemble models for high-security applications 4. Regular security audits and adversarial testing

MODEL PERFORMANCE

Baseline Model

ATTACK ANALYSIS

Overview of Evaluated Attacks

DEFENSE EVALUATION

Overview of Evaluated Defenses

RECOMMENDATIONS

Based on Evaluation Results

## Visualizations

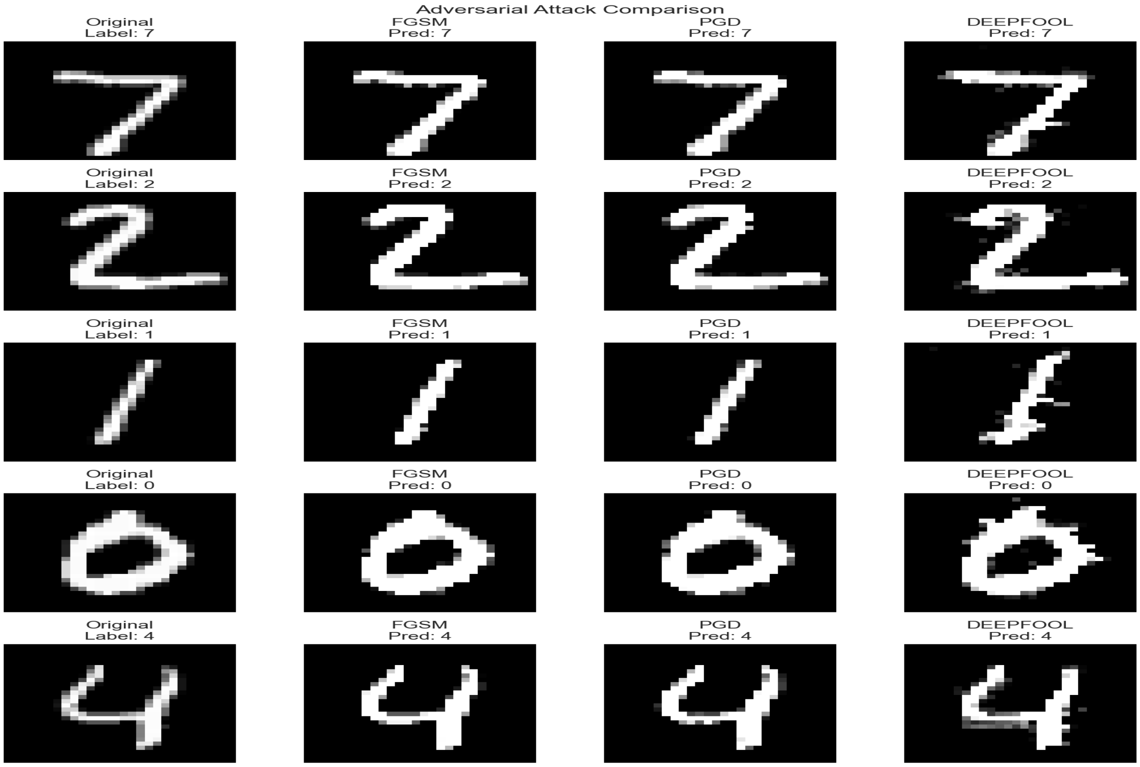

Figure: attack_comparison